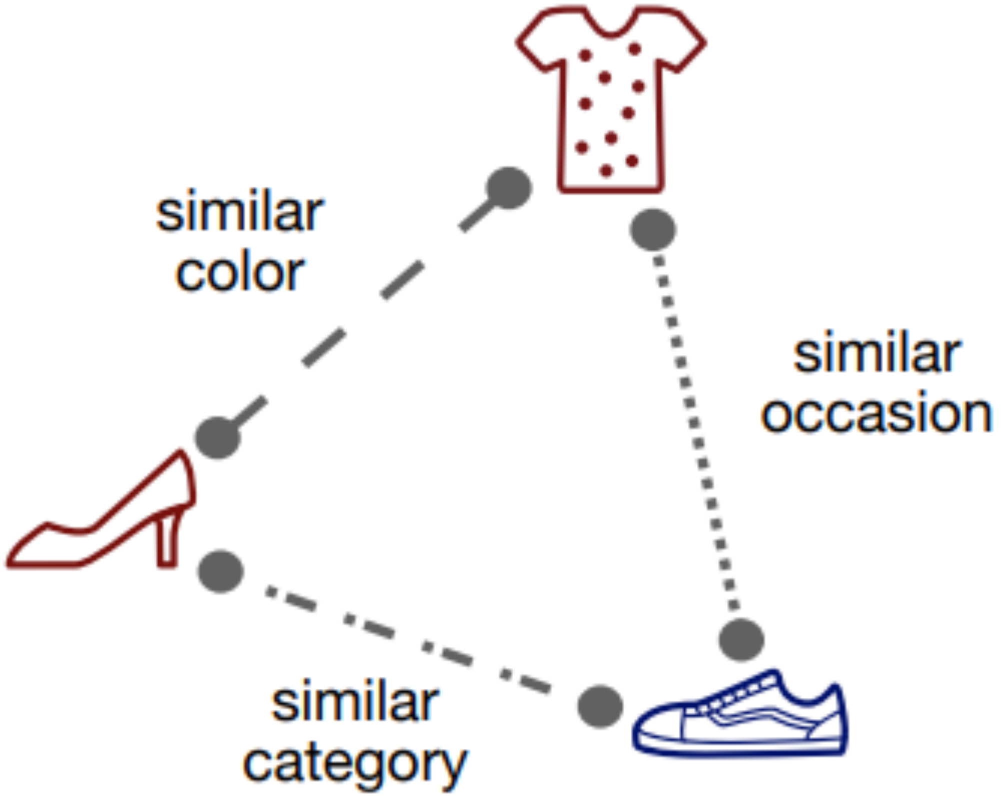
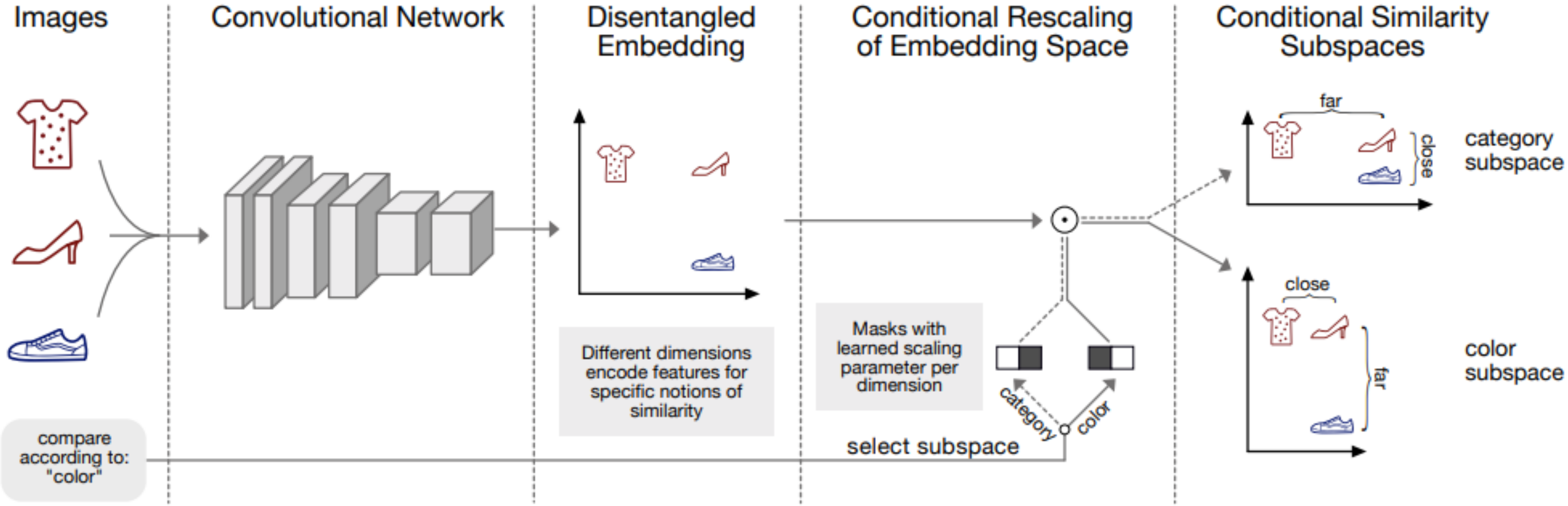
Conditional Similarity Network




Images
Convolutional Network
Disentangled Embedding
Conditional Rescaling of Embedding Space
Conditional Similarity Subspaces
Different dimensions encode features for specific notions of similarity
Masks with learned scaling parameter per dimension
category
color
select subspace
compare according to: "color"
far
close
category subspace
close
far
color subspace

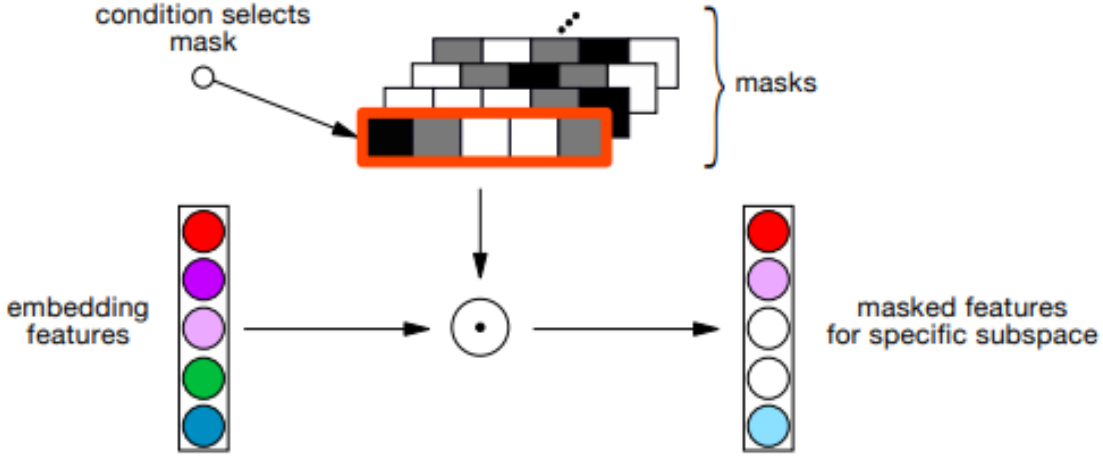
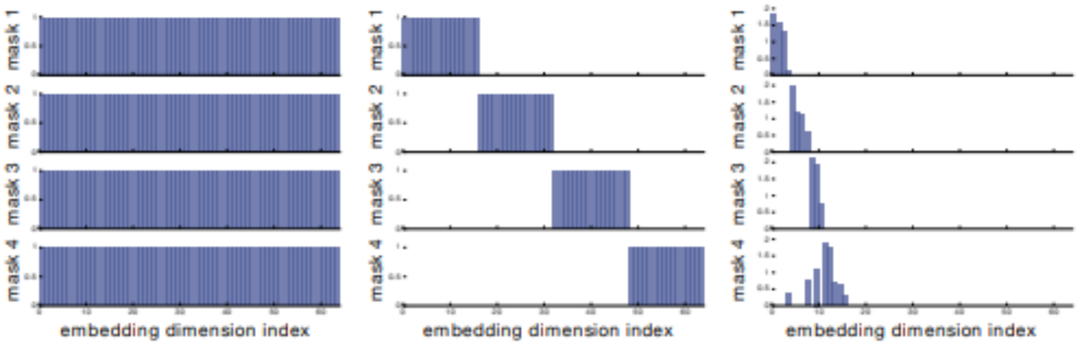


Figure 3. The masking operation selects relevant embedding dimensions, given a condition index. Masking can be seen as a soft gating function, to attend to a particular concept.



(single query) (fixed disjoint masks) (learned masks)

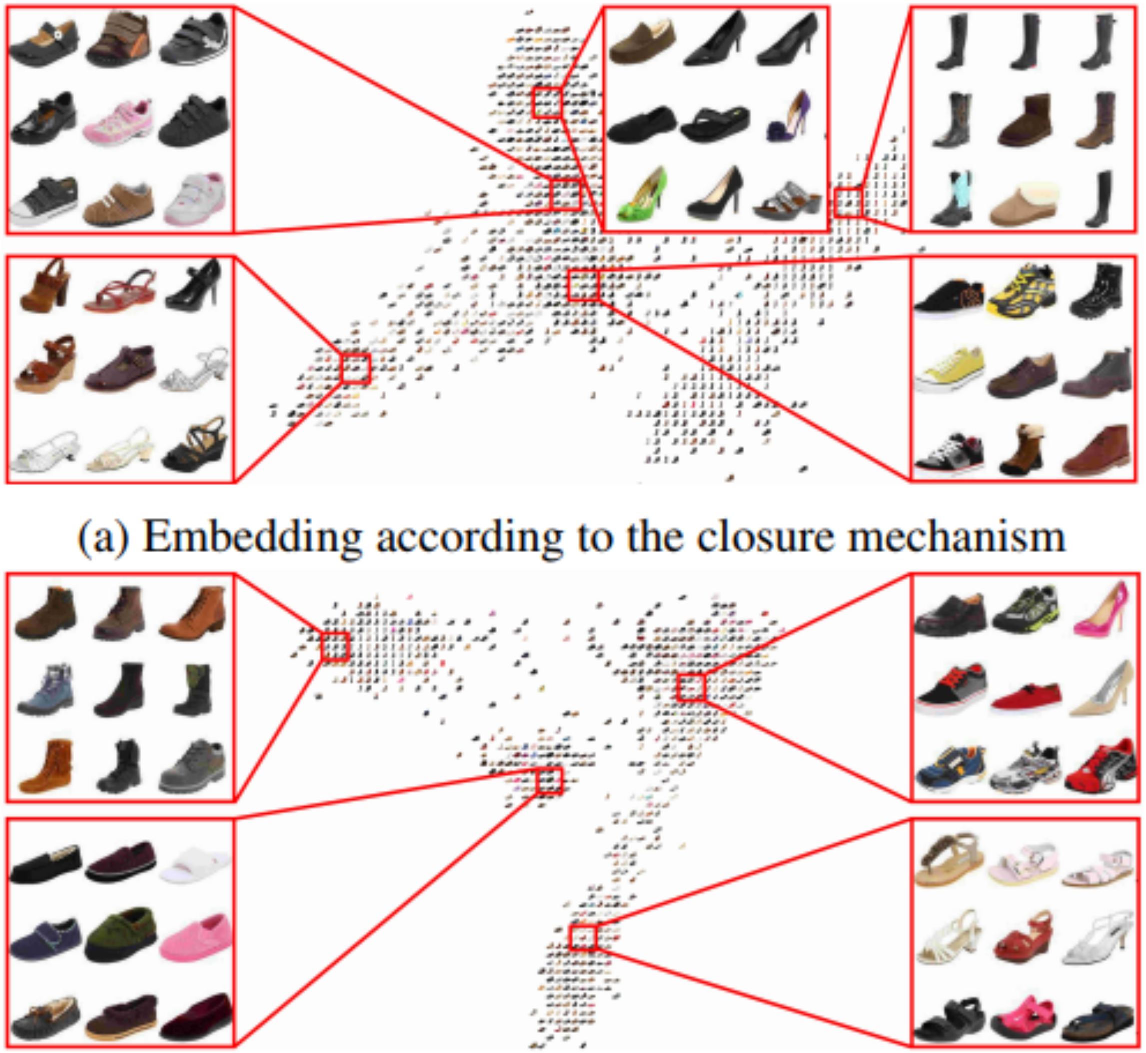
$$\mathcal{T}_\rfloor = \{(i,j,l;c) \mid s_c(x_i,x_j) > s_c(x_i,x_l)\}. \tag{1}$$

$$\begin{aligned} L_T(x_i,x_j,x_l) &= \max\{0, D(x_i,x_j) - D(x_i,x_l) + h\} \\ D(x_i,x_j) &= \|f(x_i;\theta) - f(x_j;\theta)\|_2 \end{aligned} \tag{2}$$

$$\begin{aligned} &\mathcal{L}_T(x_i,x_j,x_l,c;m,\theta) = \\ &\max\{0, D(x_i,x_j;m_c,\theta) - D(x_i,x_l;m_c,\theta) + h\} \end{aligned} \tag{4}$$

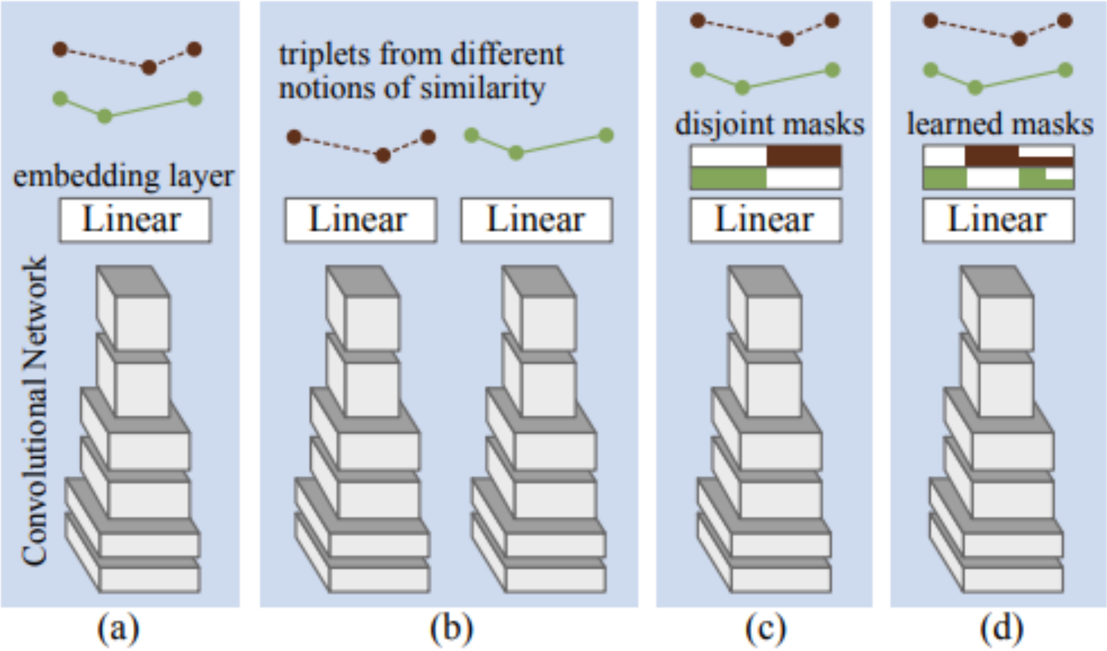
$$D(x_i,x_j;m_c,\theta) = \|f(x_i;\theta)m_c - f(x_j;\theta)m_c\|_2. \tag{3}$$

$$\begin{aligned} &\mathcal{L}_{CSN}(\mathbf{x},\{\mathbf{t},\mathbf{c}\};\mathbf{m},\theta) = \\ &\mathcal{L}_T(x_{t_0},x_{t_1},x_{t_2},c;\mathbf{m},\theta) + \lambda_1\mathcal{L}_W(\mathbf{x},\theta) + \lambda_2\mathcal{L}_M(\mathbf{m}) \end{aligned} \tag{7}$$

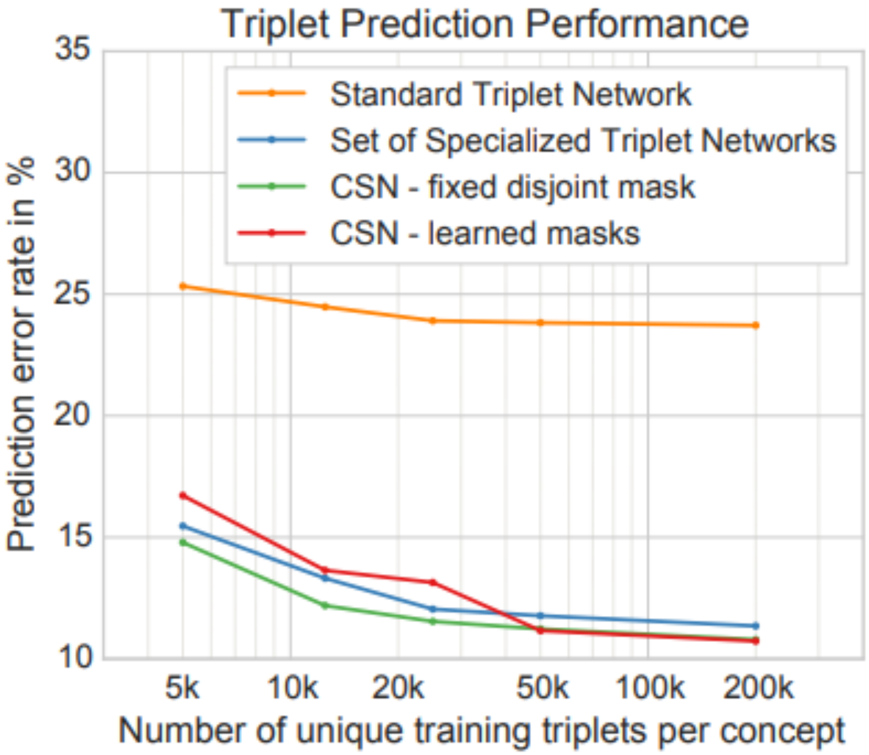


(a) Embedding according to the closure mechanism

(b) Embedding groups of boots, slippers, shoes and sandals



Method	Error Rate
Standard Triplet Network	23.72%
Set of Specialized Triplet Networks	11.35%
CSN fixed disjoint masks	10.79%
CSN learned masks	**10.73%**


Triplet Prediction Performance
Standard Triplet Network
Set of Specialized Triplet Networks
CSN - fixed disjoint mask
CSN - learned masks
Prediction error rate in %
5k
10k
20k
50k
100k
200k
Number of unique training triplets per concept